Comparison of Objective Supercell Identification Techniques Using an Idealized Cloud Model

JASON NAYLOR AND MATTHEW S. GILMORE
Department of Atmospheric Sciences, University of North Dakota, Grand Forks, North Dakota

RICHARD L. THOMPSON AND ROGER EDWARDS
Storm Prediction Center, Norman, Oklahoma

ROBERT B. WILHELMSON
Department of Atmospheric Sciences, and the National Center for Supercomputing Applications, University of Illinois at Urbana–Champaign, Urbana–Champaign, Illinois

(Manuscript received 14 August 2011, in final form 7 February 2012)

ABSTRACT

The accuracy, reliability, and skill of several objective supercell identification methods are evaluated using 113 simulations from an idealized cloud model with 1-km horizontal grid spacing. Horizontal cross sections of vorticity and radar reflectivity at both mid- and low levels were analyzed for the presence of a supercell, every 5 min of simulation time, to develop a “truth” database. Supercells were identified using well-known characteristics such as hook echoes, inflow notches, bounded weak-echo regions (BWERs), and the presence of significant vertical vorticity.

The three objective supercell identification techniques compared were the Pearson correlation (PC) using an analysis window centered on the midlevel storm updraft; a modified Pearson correlation (MPC), which calculates the PC at every point in the horizontal using a small 3 km × 3 km analysis window; and updraft helicity (UH). Results show that the UH method integrated from 2 to 5 km AGL, and using a threshold value of 180 m² s⁻², was equally as accurate as the MPC technique—averaged from 2 to 5 km AGL and using a minimum updraft threshold of 7 m s⁻¹ with a detection threshold of 0.3—in discriminating between supercells and nonsupercells for 1-km horizontal grid spacing simulations. At courser resolutions, the UH technique performed best, while the MPC technique produced the largest threat scores for higher-resolution simulations. In addition, requiring that the supercell detection thresholds last at least 20 min reduced the number of false alarms.

1. Introduction

Despite their relatively infrequent occurrence, supercells produce a large fraction of significantly severe convective events (Doswell 2001). For this reason, it is important to be able to distinguish supercells from other modes of convection both observationally and in high-resolution forecast/research models—the latter being the focus of this paper. The formal American Meteorological Society (AMS) definition of a supercell is “an often dangerous convective storm that consists primarily of a single, quasi-steady rotating updraft, which persists for a period of time much longer than it takes an air parcel to rise from the base of the updraft to the summit” (Glickman 2000, p. 745). However, that expression does not offer a quantitative means of defining a supercell.

Efforts to distinguish supercells from other modes of convection in short-term forecasts using storm-scale simulations have been performed as a part of Storm Type Operational Research Model Including Predictability Evaluation (STORMTIPE) and other spring programs at the National Severe Storms Laboratory (NSSL) since the early 1990s (e.g., Brooks et al. 1993; Wicker et al. 1997). These studies focused on using gridpoint soundings taken...
from operational models to initialize an idealized cloud model with a limited domain. Recently, it has been shown that information obtained from convective-allowing (i.e., 4-km or less horizontal grid spacing) operational models can help improve severe weather forecasts by giving insight into the general characteristics and mode of forthcoming convection (Kain et al. 2006; Weisman et al. 2008). Even though all of these studies used numerical model forecasts in some manner to predict convective mode, none offer a quantitative definition of a supercell. With the possibility of an operational warn-on-forecast system based on output from numerical models (Stensrud et al. 2009) being implemented in the near future, a rigorously tested, quantitative supercell definition could be useful for issuing warning products based on such a system.

There is also interest in the research modeling community for identifying and processing supercell characteristics for a large number of simulations (B. Jewett 2011, personal communication). High-performance computers are now capable of performing a large number of simulations in a short period of time. However, it can be time consuming and tedious to manually examine the output from many simulations for the presence of supercells. Analysis time requirements of large suites of simulations could be greatly reduced by an automated supercell detection technique.

To date, there has been no known systematic comparison of the performance among the available supercell identification algorithms. This paper describes a systematic effort in testing the performance of three automated supercell identification methods for simulated storms at 1-km horizontal grid spacing, which, while considered coarse for research, will soon be used in storm-scale forecast models. This resolution has been shown to be sufficient in capturing the basic structure of deep convection (Bryan et al. 2003).

Distinctive radar characteristics of supercells were noted long before their recognition as a separate mode of convection by Browning (1964). Newton and Katz (1958) first observed that these storms move to the right relative to the mean environmental winds. Stout and Huff (1953), Fujita (1958), and Browning and Donaldson (1963) observed hook-shaped appendages in images of radar reflectivity factor. The 4 May 1961 Geary, Oklahoma, storm (Browning and Donaldson 1963) and the 9 July 1959 Wokingham, England, hailstorm (Browning and Ludlam 1962) both were found to possess an echo-free region—now known to be coincident with a strong updraft.

The presence of these distinctive features has been used as the basis for the conceptual model of a classic supercell (e.g., Browning 1964; Marwitz 1972; Lemon and Doswell 1979). Numerous studies since have used these models in some manner as criteria for supercell identification (e.g., Lemon 1977; Brooks et al. 1994; Moller et al. 1994; Thompson 1998; Klimowski et al. 2003; Thompson et al. 2003; Bunkers et al. 2006). Some of these studies also included threshold values of Doppler radar–detected horizontal wind shear to aid in subjectively determining mesocyclone presence.

Doswell and Burgess (1993) argue that radar features are inadequate for categorizing convective storms. Reflectivity hooks and bounded weak-echo regions (BWERs) may not be evident in heavy- or low-precipitation supercells (Moller et al. 1994), rightward storm propagation may be small (Davies and Johns 1993; Moller et al. 1994), and range limitations might cause a supercell to be missed because of smoothing of the velocity signature. Since the distinctive radar-observable characteristics of a supercell (i.e., hook echoes, rightward propagation, and BWER) are related to the presence of a mesocyclone, it seems that a more direct definition of a supercell would involve a quantitative means of detecting the presence of a mesocyclone. However, as mentioned by Moller et al. (1994), a universally accepted definition of a mesocyclone does not exist.

A quantitative approach to defining a mesocyclone commonly used in numerical modeling is to examine the linear correlation between vertical velocity ($w$) and vertical vorticity ($\zeta$). However, as is the case with radar-based mesocyclone detection, a standard methodology for this technique has not been defined. Using three-dimensional models and multi-Doppler radar data, Clark (1979) found linear correlation coefficients as high as 0.4. The correlation was performed on all points over a depth of 2–10 km. Weisman and Klemp (1984) similarly calculated the linear correlation for an idealized, right-moving supercell, but only for grid points with $w > 0 \text{ m s}^{-1}$, within a 15 km $\times$ 15 km box centered on the storm and averaged over the lowest 8 km. Their results show correlation values ranging from 0.5 to 0.8. Droegemeier et al. (1993) followed Weisman and Klemp (1984) except using $w > 1 \text{ m s}^{-1}$. The analysis window was centered on the storm with a size specified to minimize influence from nearby convection (S. Lazarus 2011, personal communication).

Their results yielded correlations as high as 0.7 for supercells and 0.78 for multicells, with the distinguishing characteristic between the two storm types being the longer duration of the correlation for supercells. Knupp et al. (1998) defined a supercell as a storm with a $w-\zeta$ correlation of 0.4 or greater over $1/3$ of the storm depth that lasts for at least two updraft-parcel cycles. Their version of correlation was calculated using only points where $w \geq 3 \text{ m s}^{-1}$. None of these studies tested their respective objective threshold values using a sample size larger than a few cases.

Recently, Kain et al. (2008) and Sobash et al. (2011) presented the concept of updraft helicity (UH) as a method
for mesocyclone detection in convection-allowing operational models. UH is the local product of $w$ and $\zeta$ integrated over a specified depth [equation to be presented in section 2d(3) herein]. Kain et al. (2008) used a UH threshold of 50 m$^2$ s$^{-2}$ to define mesocyclones in their 2-km Weather Research and Forecasting Model (WRF) runs, with a smaller threshold found to be appropriate for 4-km simulations. Sobash et al. (2011) tested UH thresholds in their 4-km WRF runs with values of 34–103 m$^2$ s$^{-2}$. However, these thresholds were determined by comparing model forecasted storm structure to the number of radar-observed mesocyclone detections; thus, it was as much a test of whether the model captured the proper environmental conditions and convective mode as it was of the proper UH thresholds. This comparison was only available once per hour, and some loss of forecast skill may have resulted from the coarse temporal resolution.

The purpose of this study is to evaluate the skill of various objective supercell identification techniques by using an idealized numerical cloud model initialized with a large number of Rapid Update Cycle-2 (RUC-2) proximity soundings that were associated with supercells in nature. The modeled storms are classified based on their simulated radar reflectivity structure and presence of vertical vorticity, then compared to objective techniques to determine which approach (and threshold) has the most skill in detecting and identifying supercells. A description of the dataset and numerical model is presented in the next section, along with a description of the various objective techniques to be tested. Results are presented in section 3 and discussion and concluding remarks are found in sections 4 and 5.

2. Methodology

a. Dataset

Each idealized simulation was initialized with one of 113 RUC-2 proximity soundings associated with supercells. This dataset is a subset of that analyzed by Thompson et al. (2003, 2007). The soundings represent a bilinear interpolation between the nearest four horizontal grid points to the observing station closest to the storm. The soundings were generally within 30 and 40 miles of a convective storm that met certain subjective and objective criteria from Weather Surveillance Radar-1988 Doppler (WSR-88D) base scans, including the presence of reflectivity hook echoes, BWERs, and radial shear >0.008 s$^{-1}$. For a more detailed description of the dataset and the supercell criteria, refer to Thompson et al. (2003).

Since simulated storm structure is primarily a function of environmental CAPE and vertical wind shear (e.g., Weisman and Klemp 1982, 1984), most of the storms produced in the simulations using these soundings possess at least some supercell characteristics (i.e., midlevel rotation) at some point in their lifetime. The simulated storms produced by using the selected soundings include strong, isolated, long-lived supercells; rapidly decaying supercells with weak updrafts and lingering midlevel rotation that pose no severe threat; multicell clusters; and some quasi-linear convective systems. The analysis focused on only right-moving supercells, if present (i.e., left-moving supercells are not considered). Even though not all types of convection are represented in this study, we believe that the simulations offer a good test bed for the mesocyclone identification algorithms presented herein, because some rotation is present in nearly every case.

b. Numerical model

Simulations were performed using the Bryan cloud model (Bryan and Fritsch 2002). The model setup consisted of 1-km horizontal grid spacing and 250-m vertical grid spacing within a 120 km $\times$ 120 km $\times$ 20 km domain. A moving grid was used, with the velocity determined by the 0–6-km mean wind from the input sounding. Lateral boundary conditions were open-wave radiating and upper/lower boundaries were free slip. The upper boundary was rigid, with a damping layer extending upward from 16 km. A split time scheme was used (Klemp and Wilhelmson 1978) with a large time step of 3 s for advection and a small time step of 0.5 s. Each simulation was run for 7200 s of cloud time. Precipitation processes were represented by the single moment Lin et al. (1983)–like scheme used in Gilmore et al. (2004), with default Lin et al. parameters for hail.

Because of the presence of capping inversions in many of the RUC-2 soundings, the traditional thermal perturbation convective initiation technique was ineffective at producing sustained deep convection in many of the simulations. Because of this, an updraft nudging technique similar to that in Ziegler et al. (2010) was used. The nudging was performed for the first 1800 s of cloud time over a 10 km $\times$ 10 km $\times$ 3 km spheroid centered at $z = 1.5$ km. The $w$ field at all points inside this spheroid was accelerated toward a maximum value of 10 m s$^{-1}$, with the strongest nudging at the center of the spheroid and falling off to zero at the edges. The nudging settings were chosen such that every simulation produced a storm that lasted at least 30 min.

c. Subjective analysis

The criteria used for categorizing a storm as a supercell was designed to mimic that of Thompson et al. (2003) as closely as possible. That is, a storm was considered a supercell at a particular instant if all of the following criteria were met:
• The storm possessed one or more of the radar reflectivity characteristics typically associated with supercells (hook echo, inflow notch, BWER).

• Vertical vorticity $>0.004\, s^{-1}$ in an area likely to be coincident with updraft. By definition, the vorticity value used is half the $\Delta V$ shear value used by Thompson et al. (2003).

To perform this analysis, model history files were analyzed at 5-min intervals for each simulation. The analysis consisted of a visual inspection of horizontal plots of simulated radar reflectivity factor (computed following Smith et al. 1975) and specified contours of vertical vorticity at $z = 875$ and $z = 4875$ m. These levels were chosen because they roughly estimate the upper and lower bounds of possible radar beam heights of the data in Thompson et al. (2003). In addition, the use of multiple heights allows inferences to be made about the depth of the rotation and also the presence of a bounded weak-echo region.

At every analysis time, the storm was subjectively categorized as a supercell or nonsupercell. Averaging individual authors’ subjective categorizations created a composite analysis to be used in scoring the algorithms. A storm at a particular instant was classified as a supercell if the composite score was greater than 0.5, and a nonsupercell is otherwise. The results from the composite analysis are considered “truth” for the sake of comparisons with the objective classification techniques.

Although general guidelines were provided to coauthors regarding supercell classification, there was some freedom in interpretation. For instance, no criteria were specified regarding time scales, whether threshold rotation was required at both levels, spatial continuity between rotation at both levels, threshold values of simulated radar reflectivity factor, or the alignment of vertical vorticity relative to reflectivity structures.

**d. Objective classification techniques**

1) **PEARSON CORRELATION COEFFICIENT**

The Pearson correlation coefficient (PC) (see Wilks 2006) between vertical velocity and vertical vorticity is defined as

$$R = \frac{w' \zeta'}{\sigma_w \sigma_\zeta},$$

(1)

where $w$ is vertical velocity, $\zeta$ is the vertical component of vorticity, $\sigma$ is the standard deviation, and primed quantities indicate the deviation from the mean. This correlation generally has been calculated using only grid points with positive vertical velocity, within an analysis window selected to capture the updraft region of a particular storm. The single correlation value computed in (1) for each altitude is then averaged over some vertical depth. Previous studies have used different subregion sizes, storm-selection criteria, averaging depths, and minimum $w$ criteria for selecting grid points; and all of these choices influence the calculation.

In this study, PC was standardized to a single 9 km $\times$ 9 km analysis window centered on the maximum domain updraft at 6 km—similar to the method used in the studies cited above. This subregion size was selected because it is the approximate size of supercell updrafts from historical idealized simulations with 1-km resolution (e.g., Klemp and Wilhelmson 1978; Weisman and Klemp 1984; Rotunno and Klemp 1985). Thus, this subregion should encompass the main convective updraft of a storm and limit the likelihood of nearby weaker updrafts from influencing the correlation value. Testing was performed to determine which settings for minimum $w$ value and vertical averaging depth result in the best supercell detection as defined by the threat score. Minimum $w$ thresholds used were 3, 5, 7, and 9 m s$^{-1}$; and the vertical averaging depths tested were 1–6, 1–8, and 2–5 km. Previous studies have included levels below 1 km; however, these levels were not considered for testing herein because the presence of near-surface rotation is not essential for a storm to be classified as a supercell.

2) **MODIFIED PEARSON CORRELATION**

There are several drawbacks to the traditional application of the Pearson correlation. The first drawback is the lack of spatial representation of the mesocyclone, since only one correlation coefficient is calculated for a single storm. The second drawback is the need to define an analysis window only around the storm of interest without including other strong left-moving supercells or multicells. As mentioned above, the presence of multiple updrafts in the correlation subregion may dilute the magnitude of the correlation and scenarios such as storm splitting and storm occlusion may not produce a strong $w$-$\zeta$ correlation when considering the multipdraft storm complex as a whole. While manual placement of the analysis window has been successful in past studies, simple algorithms to automate its placement can fail to give a representative PC value when there are multiple storms in the domain.

---

1 Thompson et al. (2003) analyzed 0.5° and 1.5° WSR-88D elevation scans. Assuming storms were within 200 km of the radar location, the maximum height of the radar beam would have been approximately 5.2 km AGL.

2 Only fixed depths are used in tests of vertical averaging depth. More sophisticated techniques like that used by Knupp et al. (1998) are being tested to determine if they offer improved skill over fixed depths.
As an alternative approach to the traditional Pearson correlation, a modified Pearson correlation (MPC) was developed herein. It was calculated by applying (1) at each horizontal grid point using a smaller 3 km × 3 km subset of surrounding grid points (i.e., nine total points for the \(dx = 1\) km simulations) and then averaged over a defined vertical depth. The size of this subset was chosen so that only a portion of the updraft was considered for reasons discussed above. MPC at a particular point was set to zero if >4 (44\% of the area at 1-km grid spacing) of the points in the subset exceeded a defined \(w\) threshold. This was done in order to filter out points with little or no updraft, and to make sure that the correlation was not being computed on the very edge of the updraft. This method should ensure that the maximum possible correlation is obtained everywhere in the domain, so that the user can be more certain to extract the maximum PC value for each storm of interest without concern for placement of the analysis window. Thus, it is less sensitive to user error and potentially better for automation. The presence of a supercell was determined herein by the domain maximum value of MPC. As with the traditional PC, testing was conducted to determine the influence of \(w\) threshold and averaging depth on the reliability of this technique. Minimum \(w\) thresholds and vertical averaging depths were the same as for the PC method.

3) UH

As discussed by Kain et al. (2008), updraft helicity is the vertical component of helicity integrated over a specified depth:

\[
UH = \int_{z_1}^{z_2} w \zeta \, dz,
\]

where \(z_1\) and \(z_2\) represent the upper and lower integration limits, respectively. Equation (2) was calculated at every point in the horizontal domain. The presence of a mesocyclone was determined based on the domain maximum value of UH. Kain et al. (2008) used \(z_1\) and \(z_2\) values of 2 and 5 km, which is a shallower layer than typically has been used in other objective identification techniques. Using a deeper layer for the integration may produce better results. Thus, for a fair comparison to the PC and MPC techniques, the vertical depths of 1–6 km and 1–8 km were tested in addition to 2–5 km.

e. Verification of automated techniques

Comparisons between the subjective and algorithm results were made using a 2 × 2 contingency table (e.g., see Wilks 2006). For the purposes of this analysis, the subjective results were considered to be observational truth and the results from the automated algorithms are being verified. The accuracy of the automated techniques was evaluated based on threat score—defined as the ratio of hits to the sum of hits, misses, and false alarms. Threat score was chosen for verification because it does not consider correct nonevents; thus, high threat scores only can be achieved by minimizing both false alarms and misses. The detection threshold was systematically varied for each automated technique (0.1–1 for PC and MPC and 0–400 m\(^2\) s\(^{-2}\) for UH) and the threat score was calculated at each detection threshold. The Heidke skill score does include correct nonevents and also was tested, but is not shown because the results produced the same optimal configuration for each automated technique as the threat score analysis.

3. Results

By 1200 s, all simulations have produced a storm with a strong updraft, areas of large simulated reflectivity, and at least some vertical vorticity. To reduce the number of correct nonevents resulting from times with no storm, analysis of the cases begins at \(t = 1200\) s. A total of 2373 snapshots are obtained by analyzing each of the 113 simulations at 5-min intervals between 1200 and 7200 s. However, some of these instances are removed from consideration because the automated analysis window needed for the PC technique had moved temporarily to a different storm. After removing these occurrences, 2099 instances were left. From the subjective analysis, the dataset contains 1188 instances in which a supercell is present and 911 instances of nonsupercells. At no point in any of the simulations were multiple supercells present simultaneously. Of the instances classified as nonsupercells, 563 (62\%) are because of a lack of strong updraft in the domain (i.e., no storm).

In the upcoming sections, the following nomenclature is used to discuss a particular configuration: [technique name][minimum w threshold]w[depth]. For example, pc3w25 refers to a Pearson correlation configuration using a minimum \(w\) threshold of 3 m s\(^{-1}\) and averaged over 2–5 km.

a. Pearson correlation

Figure 1 shows that, for all configurations, the threat score is maximized using a detection threshold of 0.1, and decreases with increasing detection threshold. The small spread in threat scores at detection threshold 0.1 indicates that PC performance at this threshold is not strongly influenced by configuration (i.e., averaging depth and minimum \(w\) threshold). The spread in threat scores increases at larger detection thresholds, suggesting that configuration settings become more influential. Although the spread in threat scores increases, the general tendency of the results remains
the same. For a given $w$ threshold, threat scores are largest when using an averaging depth of 2–5 km and smallest when using an averaging depth of 1–8 km.

The PC configuration that produces the largest threat score was that using a minimum updraft value of $7 \text{ m s}^{-1}$ and averaged from 2–5 km (pc7w25) with a detection threshold of 0.1. A contingency table for this configuration, using a detection threshold of 0.1, is shown in Fig. 2. While there are 1732 instances when the subjective and automated methods agreed (1036 hits and 696 correct nonevents), there are also 367 instances (215 false alarms and 152 misses) in which the two methods disagreed. These disagreements result in a false alarm ratio of 0.1713 and a relatively low hit rate of 0.8721.

c. Updraft helicity

The UH technique produces slightly larger threat scores (about 0.78) than the optimal MPC configurations (c.f., Figs. 3 and 5). As with the previous two techniques, the configuration using the 2–5-km layer produces the largest threat scores. However, recall that for the UH technique the values are integrated over this depth, while in the PC and MPC techniques the values are averaged over the depth. As integration depth increases, larger threshold values are needed to produce a maximum in threat score (Fig. 5).
The largest threat score for depths considered here is achieved using an integration depth of 2–5 km and a detection threshold of 180 m$^2$ s$^{-2}$. Figure 6 shows that this UH configuration produces more hits and fewer misses than the optimal PC and MPC techniques, yielding the largest hit rate of the three techniques. The total number of false alarms is slightly less than with the MPC and PC techniques.

d. Comparison of automated technique performance

To demonstrate differences in how the various automated techniques perform under their optimal settings, a case was chosen for additional analysis in which the automated techniques disagreed on storm type. The case contains a storm that was subjectively classified as a supercell at all times between $t = 1200$ and $t = 7200$ s. Both the optimal MPC and UH techniques agree with the subjective classification at all times. However, the PC technique fails to detect this storm as a supercell at numerous times throughout the simulation. This case was initialized with a RUC-2 sounding from Waterloo, Iowa, on 12 May 2000. For this study, it is not necessary for the modeled storm to match the structure/evolution of the observed storm.

By $t = 1200$ s, the storm begins to develop common supercell characteristics such as a hook-shaped appendage and BWER. Values of $\zeta$ exceed 0.01 s$^{-1}$ at both $z = 875$ and $z = 4875$ m. Over the next 600 s, the storm continues to grow in size but maintains its overall structure. By $t = 1800$ s, there is evidence of storm splitting at midlevels and by $t = 3000$ s a left-moving multicell cluster is clearly distinguishable from the right-moving supercell. Up until this point, all three automated techniques classified this storm as a supercell. At $t = 3300$ s, the PC technique produces a correlation of 0.05, which is below the established optimal threshold of 0.1. The UH technique produces a maximum value of 714 m$^2$ s$^{-2}$, while the MPC technique produces a maximum value of 0.5. Both of these values are well above the optimal supercell detection thresholds established in previous sections. This trend continues until $t = 4200$ s, when the PC value again exceeds the detection threshold. To determine why the PC method stops classifying this storm as a supercell after $t = 2700$ s, this time period is analyzed in more detail.

At $t = 2700$ s, the storm possesses a strong, circular, midlevel updraft centered slightly to the northeast of the low-level hook echo (Fig. 8a). A large portion of the updraft inside the PC analysis window is coincident with positive values of $\zeta$. At this time, MPC and UH values are large over most of the updraft, and PC is 0.44. There is good spatial agreement between the UH and MPC methods as to the placement of the mesocyclone, since the contours of these values have significant overlap. Thus, it is clear why all three methods classified the storm as a supercell.

3 For this study, it is not necessary for the modeled storm to match the structure/evolution of the observed storm.
By \( t = 3900 \) s, the region of \( w > 7 \text{ m s}^{-1} \) at \( z = 4875 \) m increases along the rear flanking line, giving the overall updraft a more elongated appearance (Fig. 8b). UH values exceeding the detection threshold are located farther to the southwest, in a region coincident with the hook echo and flanking line. Inside the PC analysis window, however, there is less overlap between contours of updraft and positive \( \zeta \) than at \( t = 2700 \) s. This is likely why the PC value has decreased to \(-0.06\) at this time. A test was performed to determine if PC values would be increased by only including points with positive \( \zeta \), in addition to exceeding the minimum \( w \) threshold (not shown). This test did not improve forecast skill and actually yielded lower PC values at many times. The large 9 km × 9 km box in the PC method may have caused the storm to not be detected as a supercell. To test this theory, a 3 km × 3 km box was used (as in the MPC method), but still centered on the max updraft. This resulted in a PC value of \(-0.37\) at \( t = 3900 \) s, meaning that, for this particular case, the placement of the correlation window is at least as important as the size of the window.

e. Temporal criteria

The results presented above demonstrate the ability of the automated techniques to detect supercell characteristics at a specific point in time. However, the AMS definition of a supercell includes a stipulation that the rotating updraft be present for some period of time. The goal of the analysis presented in this section is to determine if the false alarms present in the automated techniques have shorter durations than the positive detections (hits). Because the PC technique had lower skill than the MPC and UH techniques, and because of the numerous instances in which samples were disregarded because of improper
automated placement of the PC analysis window, it is not considered in this analysis. With PC technique thus disregarded, it is no longer necessary to remove times when the PC analysis box was placed incorrectly. All times from $t = 1200–7200 \text{ s}$ for all cases are used for this subsequent analysis.

Instantaneous false alarms were grouped together based on the number of consecutive instances, with the same procedure applied to positive detections. For example, a case with six consecutive positive detections would be categorized as one “hit” that lasted for 1800 s (30 min). Detection of four consecutive false alarms would be categorized as a singular false alarm with duration of 20 min. For both the MPC and UH cases, a large proportion of the false alarms have a duration of $\geq 20 \text{ min}$ (Figs. 9b,d), while only a small proportion of the hits persist for $<20 \text{ min}$ (Figs. 9a,c). There are two scenarios that could produce short-lived false alarms: 1) a short-lived rotating updraft from a nonsupercell exists for a brief period of time, or 2) the automated technique detects a supercell shortly before it was detected subjectively. Either way, it seems from Fig. 9 that requiring that the automated algorithm to exceed the specified threshold for at least 20 min could eliminate a large proportion of false alarms. The overall improvement in skill would be greater for the UH technique, since more false alarms would be reduced and fewer hits eliminated than the MPC technique.

f. Sensitivity to horizontal grid spacing

To determine the relative performance of the three techniques at different horizontal resolutions, additional simulations were performed. A subset of 30 cases was randomly selected for simulation using $\Delta x$ of 2 km, 500 m, and 250 m. Upon performing the subjective analysis, 10 cases were discarded from the subsequent analysis—reasons for which include incorrect PC analysis window placement and nonexistent convection. Only the “optimal” configurations of the automated techniques (determined from Figs. 1, 3, and 5 and mentioned in sections 3a–c above) were tested. Results from the remaining 20 cases are shown in Fig. 10. The PC technique achieved the lowest threat score at all horizontal grid spacing, but showed limited variability in both threat score and optimal detection threshold as a function of horizontal grid spacing. Both the UH and MPC techniques produced notably larger threat scores; however, the performance of these two techniques as a function of grid spacing differed markedly. The largest threat score for the UH technique occurred in the $\Delta x = 2000 \text{ m}$ simulations, with threat scores decreasing as $\Delta x$
determined based on the $\alpha = 0.05$ level. The results from these tests show that, for the $\Delta x = 1$ km simulations, the MPC and UH techniques are both significantly better than the PC technique; however, the performance of the UH technique is not significantly better than the MPC technique (Table 1). The remaining results in Table 1 should be interpreted with some caution because of the much smaller sample size for these simulations compared to the $\Delta x = 1$ km simulations (i.e., $n = 20$ versus $n = 113$, respectively). With that in mind, note that Table 1 shows that the UH technique performs significantly better than the MPC and PC techniques in the $\Delta x = 2$ km simulations. However, when grid spacing is reduced to 250 m, the MPC technique performs significantly better than both the PC and MPC techniques.

4. Discussion

Although other combinations of depth and $w$-threshold settings could have been tested for the PC and MPC techniques, such tests may not be necessary. Both the PC and MPC tests revealed that the largest threat scores were achieved when using the smallest vertical averaging depth. Based on this finding, it is unlikely that larger depths (e.g., 2–10 km) will result in larger threat scores. Smaller depths would be undesirable because the presence of shallow updraft rotation could increase the number of false alarms. Tests with $w$ thresholds $<3$ m s$^{-1}$ and $>9$ m s$^{-1}$ probably are unnecessary also, since the PC results show small spread in the threat scores at the peak correlation threshold value (i.e., 0.1) and the MPC technique appears to be more dependent on averaging depth than minimum $w$ threshold.

For all three automated techniques, the largest threat score was achieved using a depth of 2–5 km. Threat scores decreased as depth increased. The physical reason for this finding may differ between the UH and correlation techniques. The UH technique could be particularly sensitive to updraft tilting in strongly sheared environments because it is vertically integrated at a specific $x$, $y$ location. When larger depths are used for UH calculations on storms with strongly tilted updrafts, the updraft may tilt out of the column at the specific $x$, $y$ grid point, yielding zero values at certain levels and weakening

Table 1. Result from one-sided matched pair $t$ tests between two specified techniques for all simulations. The two techniques being compared are shown in the top row. The left number in each cell is the $t$ statistic, with the $p$ value to the right in parentheses. Values in bold represent differences that are statistically significant at the 0.05 level.

<table>
<thead>
<tr>
<th>Resolution</th>
<th>PC – MPC = 0</th>
<th>MPC – UH = 0</th>
<th>PC – UH = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000 m</td>
<td>-1.29 (0.0956)</td>
<td>-2.86 (0.0014)</td>
<td>-2.51 (0.0046)</td>
</tr>
<tr>
<td>1000 m</td>
<td>-6.47 ($6.6 \times 10^{-15}$)</td>
<td>-0.39 (0.3472)</td>
<td>-6.54 ($3.9 \times 10^{-17}$)</td>
</tr>
<tr>
<td>500 m</td>
<td>-1.83 (0.0306)</td>
<td>0.96 (0.1656)</td>
<td>0.2894 (0.3861)</td>
</tr>
<tr>
<td>250 m</td>
<td>-3.51 ($9.78 \times 10^{-5}$)</td>
<td>1.66 (0.0455)</td>
<td>-1.31 (0.0917)</td>
</tr>
</tbody>
</table>

FIG. 10. Threat score as a function of horizontal grid spacing for the UH (dashed line with triangles), MPC (dotted line with squares), and PC (dashed-dotted line with circles) techniques. Numbers in parentheses represent the detection threshold that resulted in the largest threat score for a particular technique. Values of UH are in m$^2$ s$^{-2}$ while PC and MPC values are dimensionless.
the overall value of UH. For both MPC and PC calculations, since the technique typically used is to average all correlation values over the depth, the lack of updraft rotation at a particular level (due to either absence of updraft or tilting outside of the analysis window) also would weaken the overall correlation coefficient. As a result, shallow supercells may go undetected when large depths are used. In a practical setting, this would include tropical-cyclone supercells, which tend to be relatively shallow (e.g., McCaul 1991; McCaul and Weisman 1996); such storms specifically were excluded from the Thompson et al. (2003) dataset.

All three automated techniques produced many more false alarms than misses. It is possible that some of these false alarms are the result of subjective misclassification due to limitations in methodology of the subjective identification technique. By only considering two vertical levels in the subjective classification, it is possible that some of the subjective criteria were absent at these levels, yet present at some vertical level(s) between them. In this scenario, a storm would be subjectively classified as a nonsupercell, yet detected by the automated technique as a supercell—yielding a false alarm. However, since all three automated techniques produced the largest threat score with the same vertical depth (2–5 km), and all produced approximately the same number of false alarms, it is unlikely that this effect influenced one particular automated technique more than another.

The results demonstrate that the commonly used Pearson correlation produced the lowest threat scores of the three automated techniques. However, this is not because linear correlation is a poor method for detecting supercells. If this were true, the MPC technique should have produced similarly small threat scores. Rather, the PC technique is hindered by the need to define an analysis window around the storm of interest. Additionally, it was found that MPC technique produced the largest threat scores of the three techniques for the higher-resolution simulations (i.e., $\Delta x$ of 500 and 250 m) while the UH technique produced the largest threat scores for lower-resolution simulations (e.g., $\Delta x = 2$ km). These performance differences among the techniques were statistically significant and reasons for these differences are a topic for future research.

The reader may wonder about the robustness of the subjective analysis results. It was suggested by one reviewer that if a different group of contributors were to perform the subjective analysis, the results would likely vary somewhat. The contributors to the supercell subjective analysis independently agreed on 77% of the 2099 analysis snapshots obtained from the simulations with 1-km horizontal resolution. Most of the disagreement occurred near the end of the simulations when many of the supercells were decaying.

To examine any potential variability in the results, the threat scores shown in Fig. 10 were recomputed using only the subjective analysis results from one analyst. The results from this test showed that the threat scores varied by 1% or less from the values shown in Fig. 10—a value that we feel is not significant.

Finally, the model results in this study are dominated by the presence of supercells. Less than 25% of the 2099 analysis snapshots contain strong convection of a mode other than supercells. Further testing is needed to better determine the false alarm rates from other forms of convection, such as squall lines, multicell systems, and ordinary cells.

5. Conclusions

This study evaluated the accuracy, reliability, and skill of several automated mesocyclone detection algorithms using a dataset of 113 idealized simulations at 1-km horizontal grid spacing by comparing the automated results to the results obtained from a subjective classification of the storms produced in the simulations. The goals were to test the sensitivity of each technique to various configuration settings and determine which technique and configuration detected simulated supercells with the greatest skill. The following conclusions were found:

- The PC, MPC, and UH techniques all produced the largest threat score when the analysis was performed over a depth of 2–5 km. Threat scores decreased as vertical depth increased.
- The largest threat scores for both the PC and MPC techniques were achieved when using all points with $w > 7$ m s$^{-1}$ and averaged from $z = 2$–5 km. Compared to previous studies that have used the PC technique, the updraft threshold is larger than typically used in the past, while the averaging depth is shallower. Additionally, the PC performed worse than the other two techniques with the fewest number of hits and largest number of misses.
- The UH technique integrated from 2 to 5 km and using a detection threshold of 180 m$^2$ s$^{-2}$ produced the largest threat score of the three techniques for the $\Delta x = 1$ km simulations; however, differences between the UH and MPC techniques were not statistically significant. At $\Delta x = 250$ m, and using a subset of cases, the MPC technique (averaged from 2–5 km and using a $w$ threshold of 7 m s$^{-1}$) produced the largest threat scores, which were found to be significantly larger than the UH technique.
- False alarms tended to have a much shorter duration than positive detections for the MPC and UH techniques (the only two tested). The majority of the false
alarms lasted less than 20 min, while the majority of the positive detections lasted more than 20 min. Based on this finding, adding a requirement that the automated detection threshold must be exceeded for >20 min should substantially reduce false alarms without drastically decreasing positive detections.

Therefore, we conclude that for simulations with $\Delta x = 1$ km, either the UH or MPC techniques could be used with nearly equal skill in automated detection of nontropical supercells in both large idealized parameter studies and in convective resolving forecast models. Additionally, the UH technique has the most skill in automatically detecting supercells in simulations with $\Delta x = 2$ km, while the MPC technique has the most skill for simulations with $\Delta x = 500$ or 250 m.

Acknowledgments. This work was supported by NSF Grant AGS-0843269. Computational resources were provided by Teragrid allocations TG-ATM100048 and TG-MCA94P023. Mr. Lawrence Burket also helped provide subjective supercell scoring of each case, and Mr. Jon Siwek helped prepare the simulations. We would also like to thank the three anonymous reviewers whose comments and suggestions improved an earlier version of this manuscript.

REFERENCES


